EM375 MECHANICAL ENGINEERING EXPERIMENTATION
TESTS BETWEEN TWO POPULATIONS AND SAMPLES
"Are Two Samples Drawn From The Same Population?"

Assume we take a set of measurements of something. At a later time, we take another set of measurements. We may need to ask questions like, "Has a manufacturing process changed between the two times I took the measurements?" In the measurements lab you measured valve balls and asked the question, "Does the seam in the ball change the diameter?" To find out, we will assume the two samples are from same population, and check (to a chosen level of confidence) whether the difference between the measured means is close to zero.

Let us define the statistics for the two samples (A and B) as:

$$y_A : \text{ mean} = \bar{y}_A \text{ ; variance} = S_A^2$$
$$y_B : \text{ mean} = \bar{y}_B \text{ ; variance} = S_B^2$$

If the process we are measuring has not changed, we should expect the mean values to be the same, or at least very "close" to each other. We need to decide how close is close. We recall the standard deviation of the mean, or standard error, and use a similar approach. We address the question by forming a new distribution, called the distribution of the "differences in means". Call this Y_{DIFF} with:

$$Y_{DIFF} = \bar{y}_A - \bar{y}_B$$

We can examine the distribution of Y_{DIFF} to see whether $Y_{DIFF} = 0$ falls inside or outside the desired confidence interval. First we find the variance of the distribution for the variable Y_{DIFF}. This variance is obtained by using the principle of pooled variance. That is, the variance of Y_{DIFF} is the pooled variance of y_A and y_B. Call this S_P^2 with:

$$S_P^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{(n_A - 1) + (n_B - 1)}$$

The standard deviation of the mean of the differences is denoted S_{DIFF}. This is comparable to the deviation of the mean for a single-variable data set.

$$S_{DIFF} = S_P \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

The confidence interval for Y is then expressed as follows:

$$\text{Confidence Interval} = \pm z.S_{DIFF} \qquad \text{for } (n_A + n_B) / 2 > 30$$

or

$$\text{Confidence Interval} = \pm t.S_{DIFF} \qquad \text{for } (n_A + n_B) / 2 < 30$$

<u>EXAMPLE:</u> A steel making process operates continuously to produce steel with a 1.20% carbon content. The content is tested daily to ensure conformity with the desired specifications. The daily test takes 30 samples, and their mean and standard deviation are found. On day 1, the mean carbon content was 1.20% with a standard deviation of 0.010. On day 2, the mean was 1.21% with a standard deviation of 0.012. Is this difference in mean values due only to chance, or can we assume something has changed in the manufacturing process?

<u>SOLUTION:</u> The two sample distributions are described by:

$$\overline{C_A} = 1.20 \; ; \quad S_A = 0.010$$
$$\overline{C_B} = 1.21 \; ; \quad S_B = 0.012$$

The mean of the distribution of differences is (1.21) - (1.20) = 0.01, with a deviation, S_P, based on the pooled variance of the sample distributions, C_A and C_B.

$$S_P = \sqrt{\frac{(30-1)\times 0.01^2 + (30-1)\times 0.012^2}{30+30-2}} = 0.01104$$

The deviation of the difference, S_{DIFF}, is thus:

$$S_{DIFF} = 0.01104\sqrt{\frac{1}{30}+\frac{1}{30}} = 0.00285$$

We choose to use a 95% confidence interval. We have (n1 + n2)/2 $\geq$30, so we use the normal distribution. Using the normal distribution table, we look up the z-value that equates to the (half) area under the probability curve being 0.95 / 2 = 0.4750, and obtain z = 1.96. The 95% confidence interval is:

$$95\% \text{ Confidence Interval} = \pm z.S_{DIFF} = \pm 1.96 \times 0.00285 = \pm 0.00559$$

We now have to decide what these statistics mean. We are asking if the two sets of measurements could have come from the same population. If so, we are basically asking if the observed difference in the averages is “statistically zero”. In other words, is the measured difference of the means (0.01) <u>inside</u> the range ±0.00559? The answer is “No”. This means that (at the 95% confidence limit) <u>the difference in observations is not due to chance alone, and the two samples come from different populations.</u> The process has changed.

The previous procedure enabled us to determine (with a yes/no answer) whether two samples were drawn from the same population. Alternatively, we may have been asked to give a level of confidence that the process had not changed. We do this by calculating the z-statistic for $Y_{DIFF} = 0$.

$$Z_{Y=0} = \frac{0.00 - Y_{DIFF}}{S_{DIFF}} = \frac{-0.01}{0.00285} = -3.51$$

Thus the observed difference in mean values is 3.51 standard deviations away from zero. To find the probability that this will happen, we again turn to the normal distribution table, and look up the half-area for z=3.51. This area is 0.4998. The confidence of the result being within ±3.51 standard deviations of the mean is thus 2×0.4998 = 0.9996 = 99.96%. Thus, if we repeated the experiment lots of times, and the two samples really were from the same population, 99.96% of the time the results would be “better” than we observed this time.

This means that if our two sets of observations were from the same population (i.e. the process did not change) we would only obtain our results 0.04% of the time. We are 99.96% certain the process has changed.

Note: If the degrees of freedom are less than 30, the Student-t statistic should be used rather than the z-statistic. Statisticians have evolved an algebraically complex formula for calculating degrees of freedom. For our purposes, we will assume the degrees of freedom for the difference distribution is:

$$\text{D.O.F.} = \frac{(n_1 + n_2)}{2}$$

and use the Student-t statistic when D.O.F is less than 30.